Hands-On: Running Local LLMs with Docker Model Runner

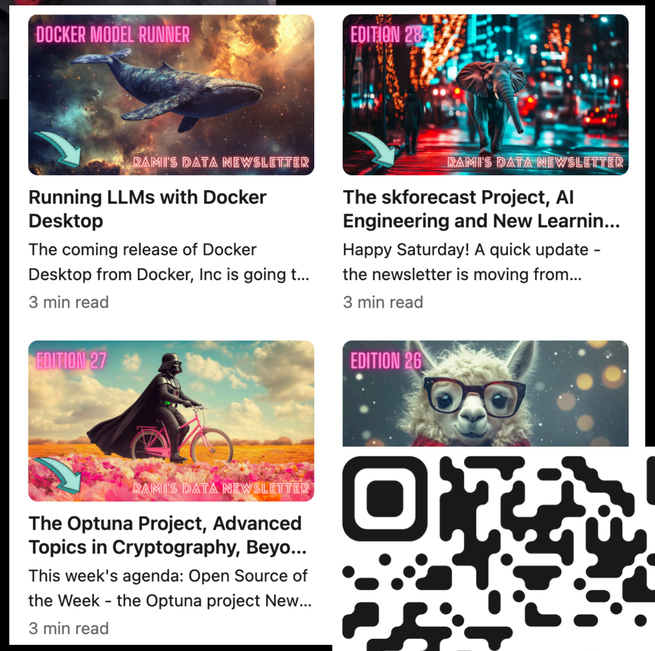
Packt Publication Workshop

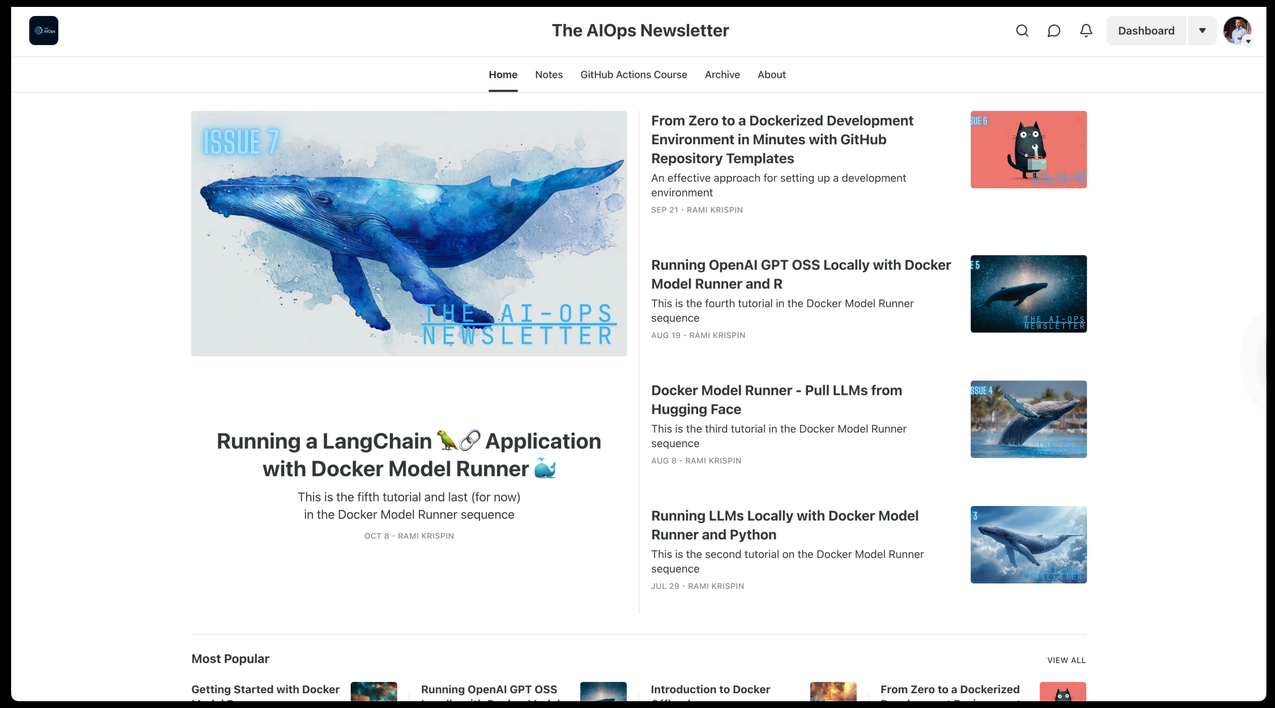
Rami Krispin, February 15th, 2026

Rami Krispin

Senior Manager - Data Science & Engineering

Author | Docker Captain 🐋 | LinkedIn Instractor





Agenda

- Introduction to DMR
- Setup
- Pull LLMs from Docker Hub & Hugging Face
- Running LLMs via the CLI
- Running LLMs with Python
- OpenCode

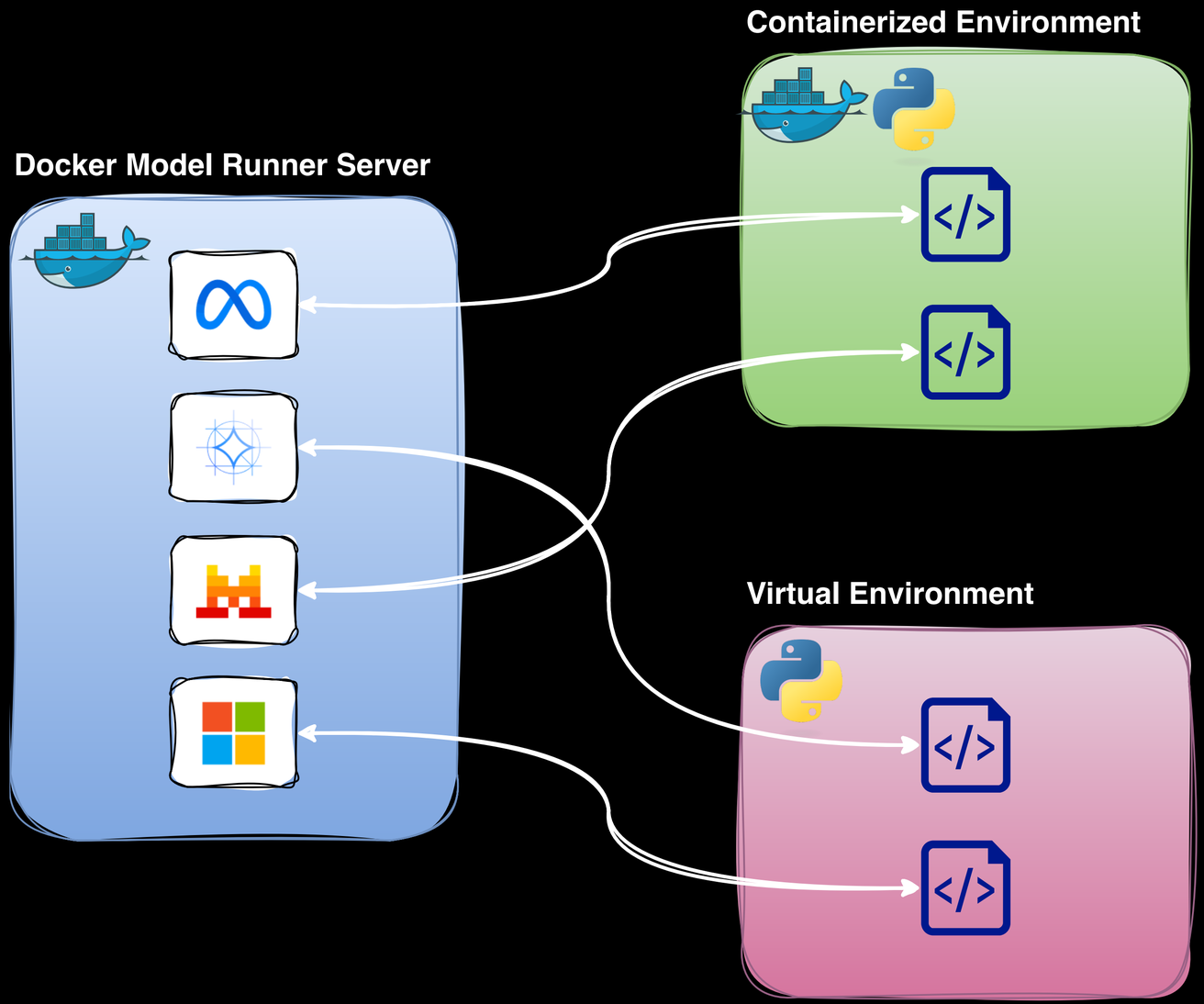
Poll

- Docker
- Python
- OpenAI Python SDK

https://github.com/RamiKrispin/dmr-workshop-packt

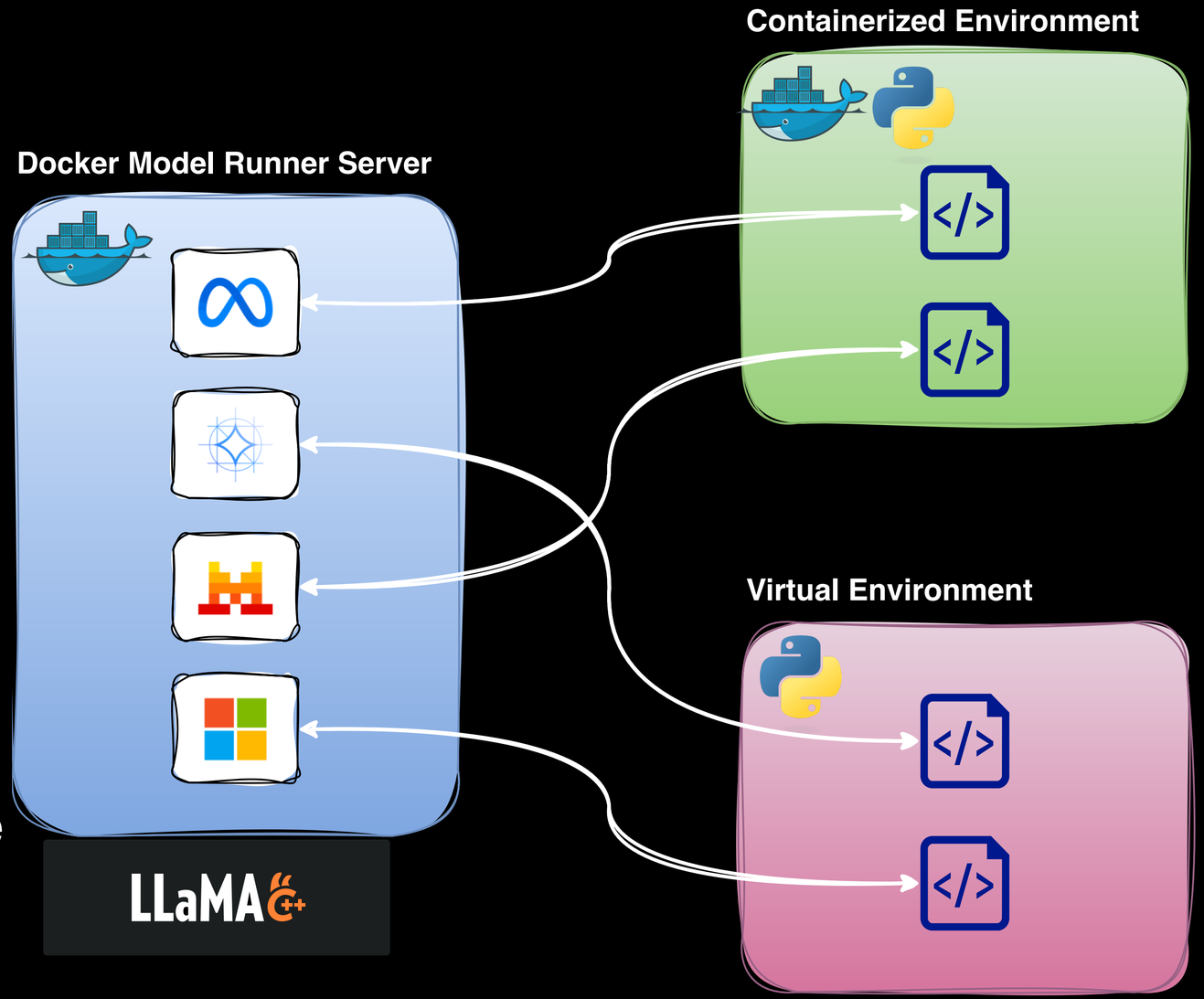
DMR Doesn’t Require Prior Docker Knowledge

Introduction to DMR



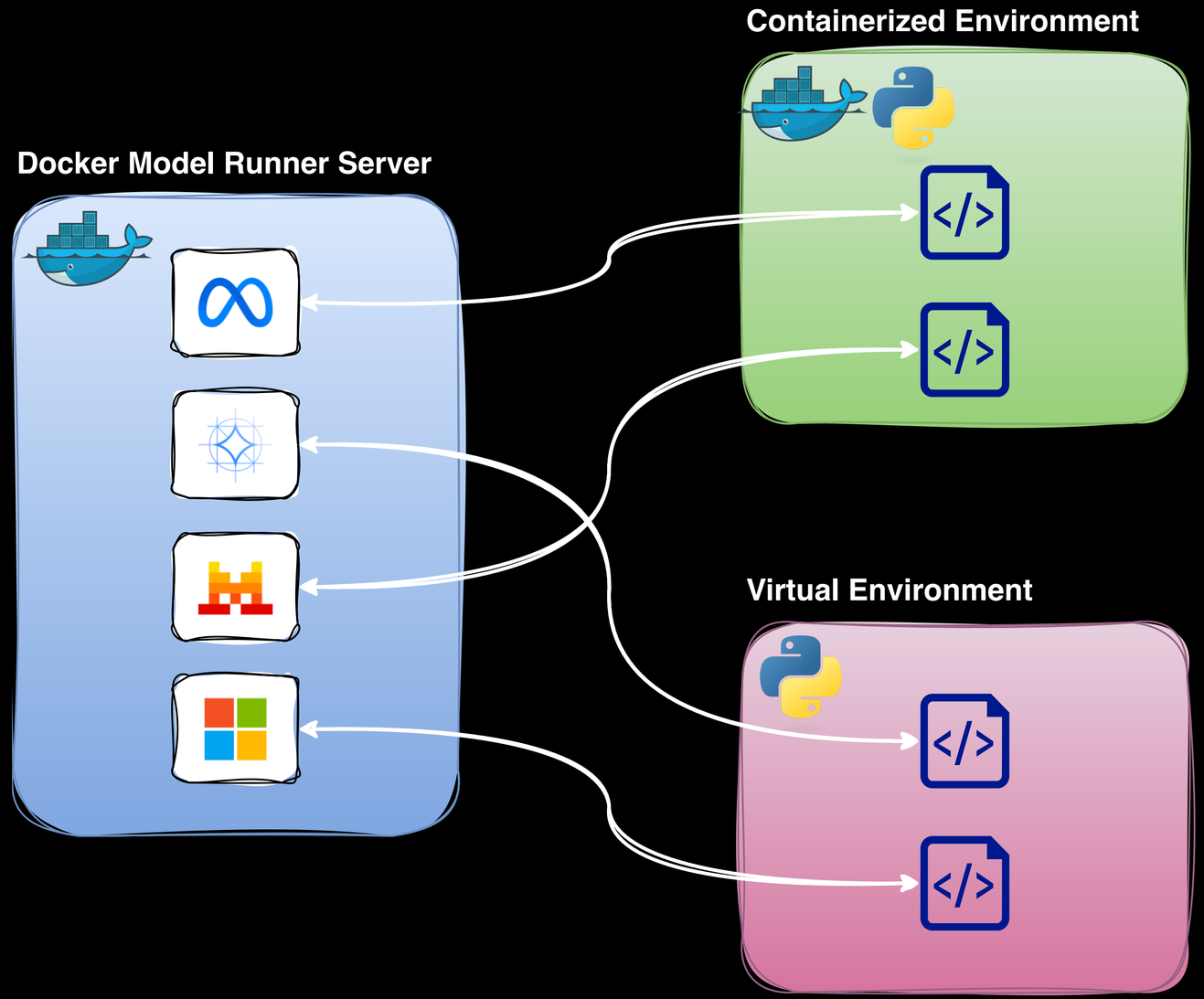
Introduction to DMR

- Docker Desktop feature
- Based on Llama.cpp
- Easy to setup and use
- Open Container Initiative
- Supports GGUF format
- OpenAI API SDK compatible



Applications

- Local development of AI applications
- Running local AI applications (local AI agents, code generation, etc.)
- Privacy



DMR Backend

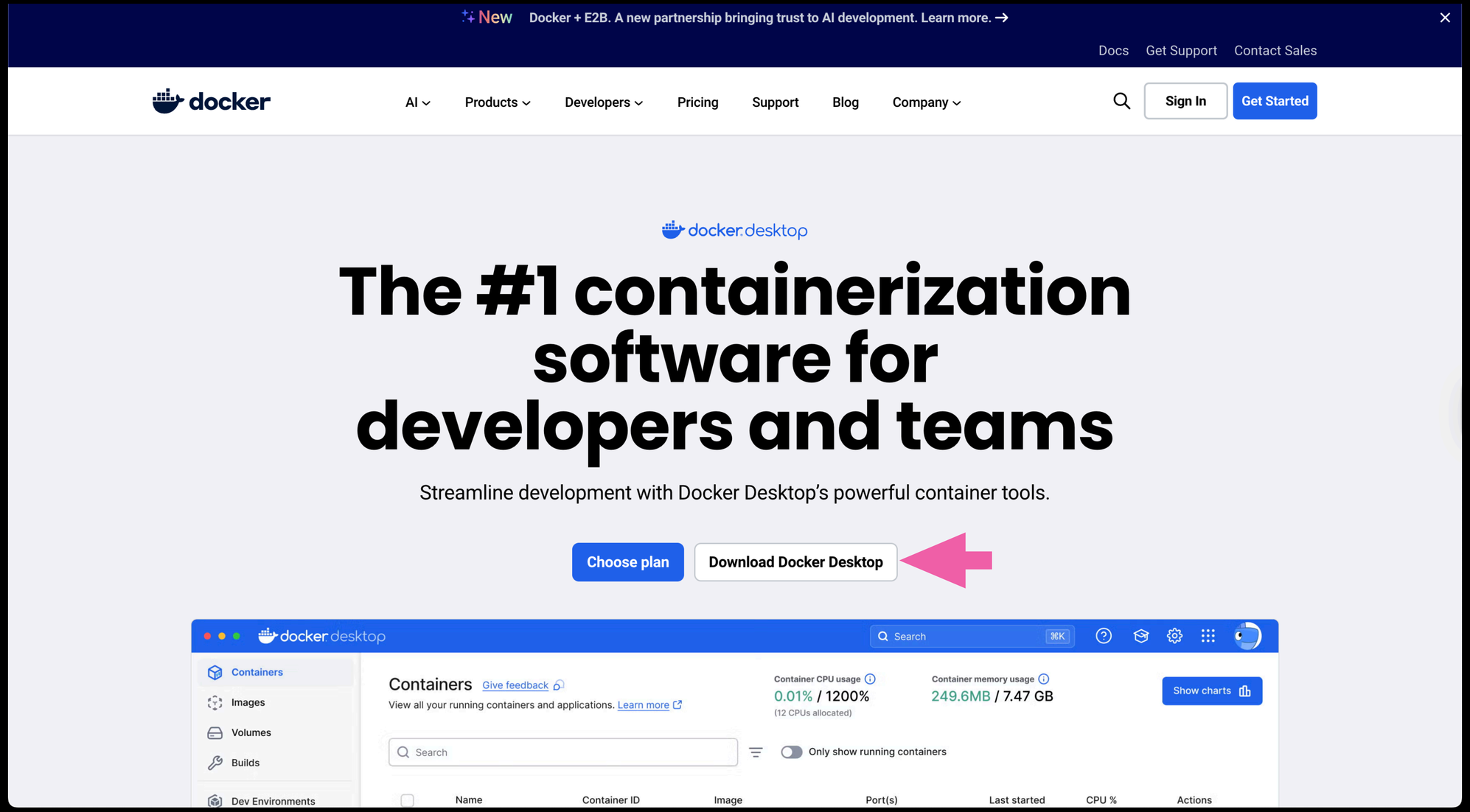

AI GUIDE TO THE GALAXY
Running
Local LLMs
with Docker
Model Runner
docker

General Requirements

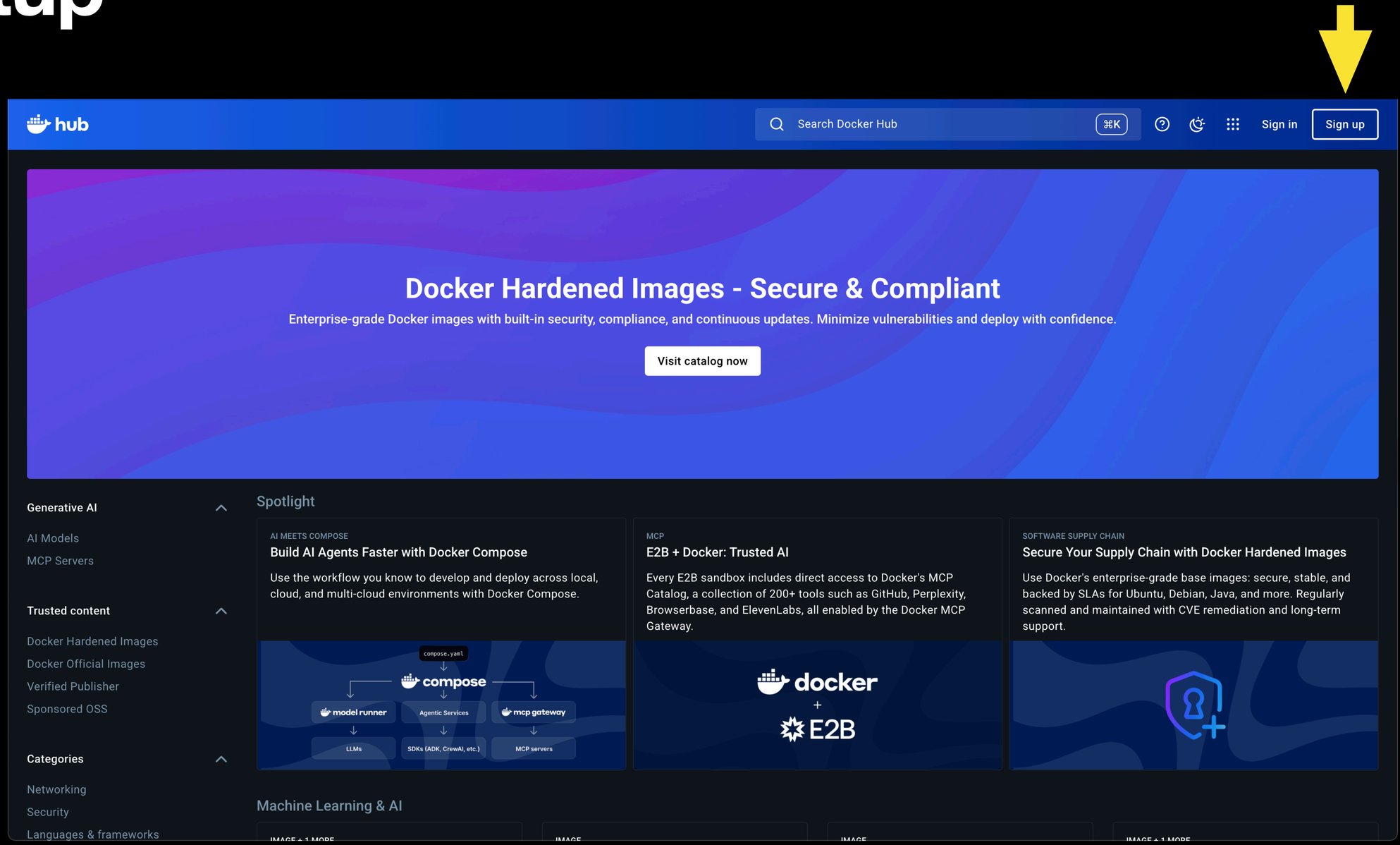
- Docker Engine (Linux)
- Docker Desktop 4.40 (Mac) or 4.41 (Windows)
- Docker Hub account
- Hugging Face API key (optional)
- OpenCode/Claude Code (optional)
- Local computational resources

Setup

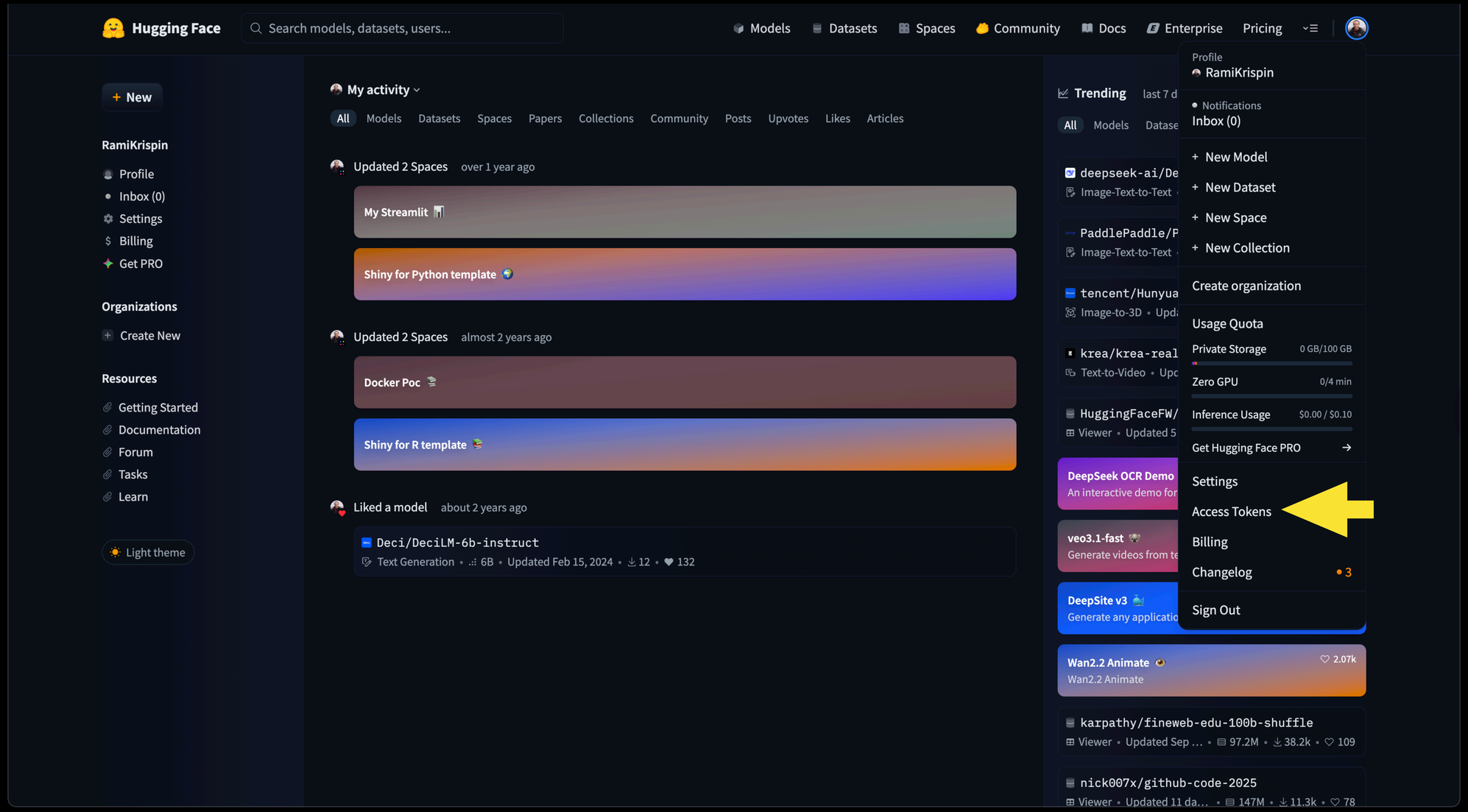
Setup


New Docker + E2B. A new partnership bringing trust to AI development. Learn more.
Docs
Get Support
Contact Sales
docker
AI
Products
Developers
Pricing
Support
Blog
Company
Sign In
Get Started
docker desktop
The #1 containerization software for developers and teams
Streamline development with Docker Desktop's powerful container tools.
Choose plan
Download Docker Desktop
docker desktop
Search ⌘K
Containers
Images
Volumes
Builds
Dev Environments
Containers
Give feedback
View all your running containers and applications. Learn more
Container CPU usage
0.01% / 1200%
(12 CPUs allocated)
Container memory usage
249.6MB / 7.47 GB
Show charts
Search
Only show running containers
Name
Container ID
Image
Port(s)
Last started
CPU %
Actions

Setup


hub
Search Docker Hub
⌘K
Sign in
Sign up
Docker Hardened Images - Secure & Compliant
Enterprise-grade Docker images with built-in security, compliance, and continuous updates. Minimize vulnerabilities and deploy with confidence.
Visit catalog now
Generative AI
AI Models
MCP Servers
Trusted content
Docker Hardened Images
Docker Official Images
Verified Publisher
Sponsored OSS
Categories
Networking
Security
Languages & frameworks
Spotlight
AI MEETS COMPOSE
Build AI Agents Faster with Docker Compose
Use the workflow you know to develop and deploy across local, cloud, and multi-cloud environments with Docker Compose.
compose.yaml
compose
model runner
Agentic Services
mcp gateway
LLMs
SDKs (ADK, CrewAI, etc.)
MCP servers
MCP
E2B + Docker: Trusted AI
Every E2B sandbox includes direct access to Docker's MCP Catalog, a collection of 200+ tools such as GitHub, Perplexity, Browserbase, and ElevenLabs, all enabled by the Docker MCP Gateway.
docker
+
E2B
SOFTWARE SUPPLY CHAIN
Secure Your Supply Chain with Docker Hardened Images
Use Docker's enterprise-grade base images: secure, stable, and backed by SLAs for Ubuntu, Debian, Java, and more. Regularly scanned and maintained with CVE remediation and long-term support.
Machine Learning & AI

Setup


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Community
Docs
Enterprise
Pricing
+ New
RamiKrispin
Profile
Inbox (0)
Settings
Billing
Get PRO
Organizations
Create New
Resources
Getting Started
Documentation
Forum
Tasks
Learn
Light theme
My activity
All
Models
Datasets
Spaces
Papers
Collections
Community
Posts
Upvotes
Likes
Articles
Updated 2 Spaces over 1 year ago
My Streamlit
Shiny for Python template
Updated 2 Spaces almost 2 years ago
Docker Poc
Shiny for R template
Liked a model about 2 years ago
Deci/DeciLM-6b-instruct
Text Generation • 6B • Updated Feb 15, 2024 • 12 • 132
Trending
last 7 d
All
Models
Datase
deepseek-ai/De
Image-Text-to-Text
PaddlePaddle/P
Image-Text-to-Text
tencent/Hunyua
Image-to-3D • Upd
krea/krea-real
Text-to-Video • Upc
HuggingFaceFW/
Viewer • Updated 5
DeepSeek OCR Demo
An interactive demo for
veo3.1-fast
Generate videos from te
DeepSite v3
Generate any applicatio
Wan2.2 Animate
2.07k
Wan2.2 Animate
karpathy/fineweb-edu-100b-shuffle
Viewer • Updated Sep … • 97.2M • 38.2k • 109
nick007x/github-code-2025
Viewer • Updated 11 da… • 147M • 11.3k • 78
Profile
RamiKrispin
Notifications
Inbox (0)
+ New Model
+ New Dataset
+ New Space
+ New Collection
Create organization
Usage Quota
Private Storage 0 GB/100 GB
Zero GPU 0/4 min
Inference Usage $0.00 / $0.10
Get Hugging Face PRO
Settings
Access Tokens
Billing
Changelog • 3
Sign Out

Setup

opencode

Install

The easiest way to install OpenCode is through the install script.

```
curl -fsSL https://opencode.ai/install | bash
```

You can also install it with the following commands:

Using Node.js

NPM BUN PNPM YARN

```
npm install -g opencode-ai
```

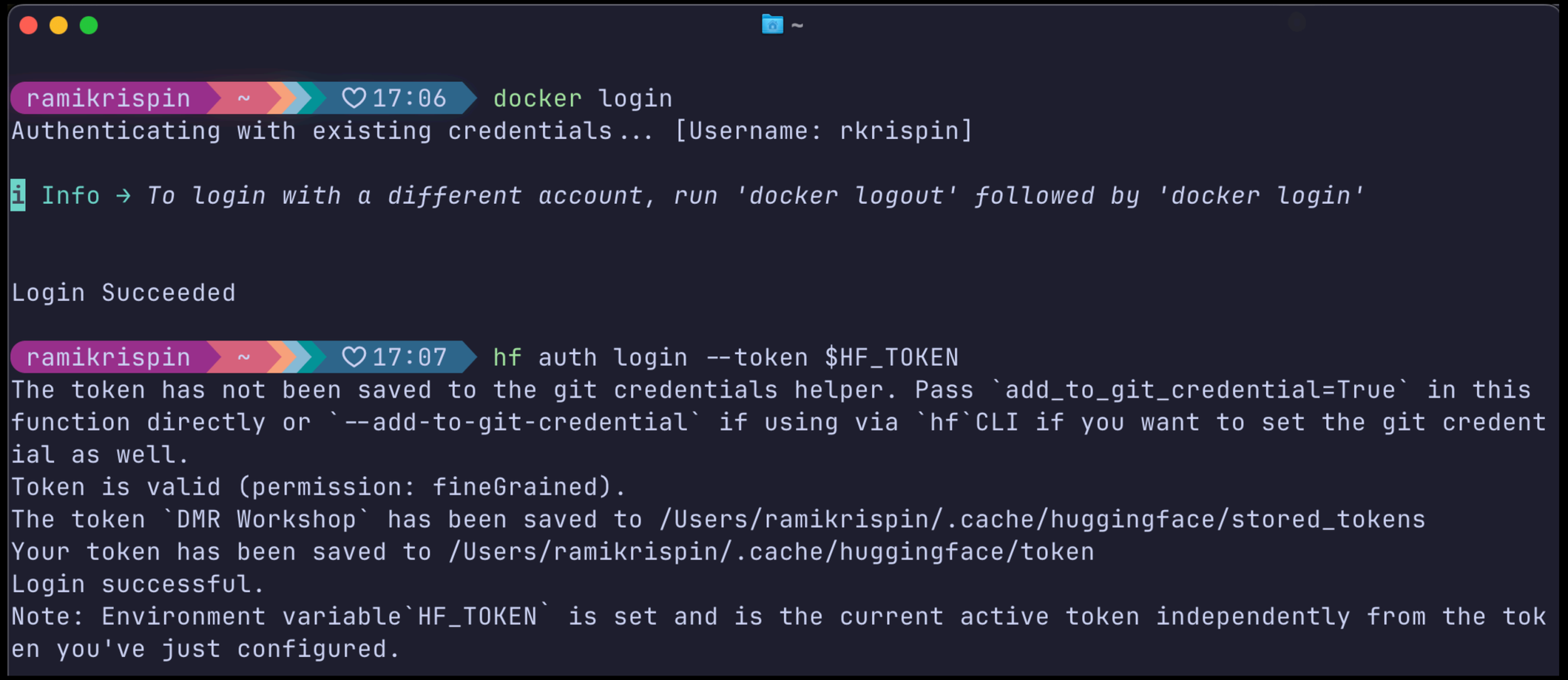
Using Homebrew on macOS and Linux

```
brew install sst/tap/opencode
```

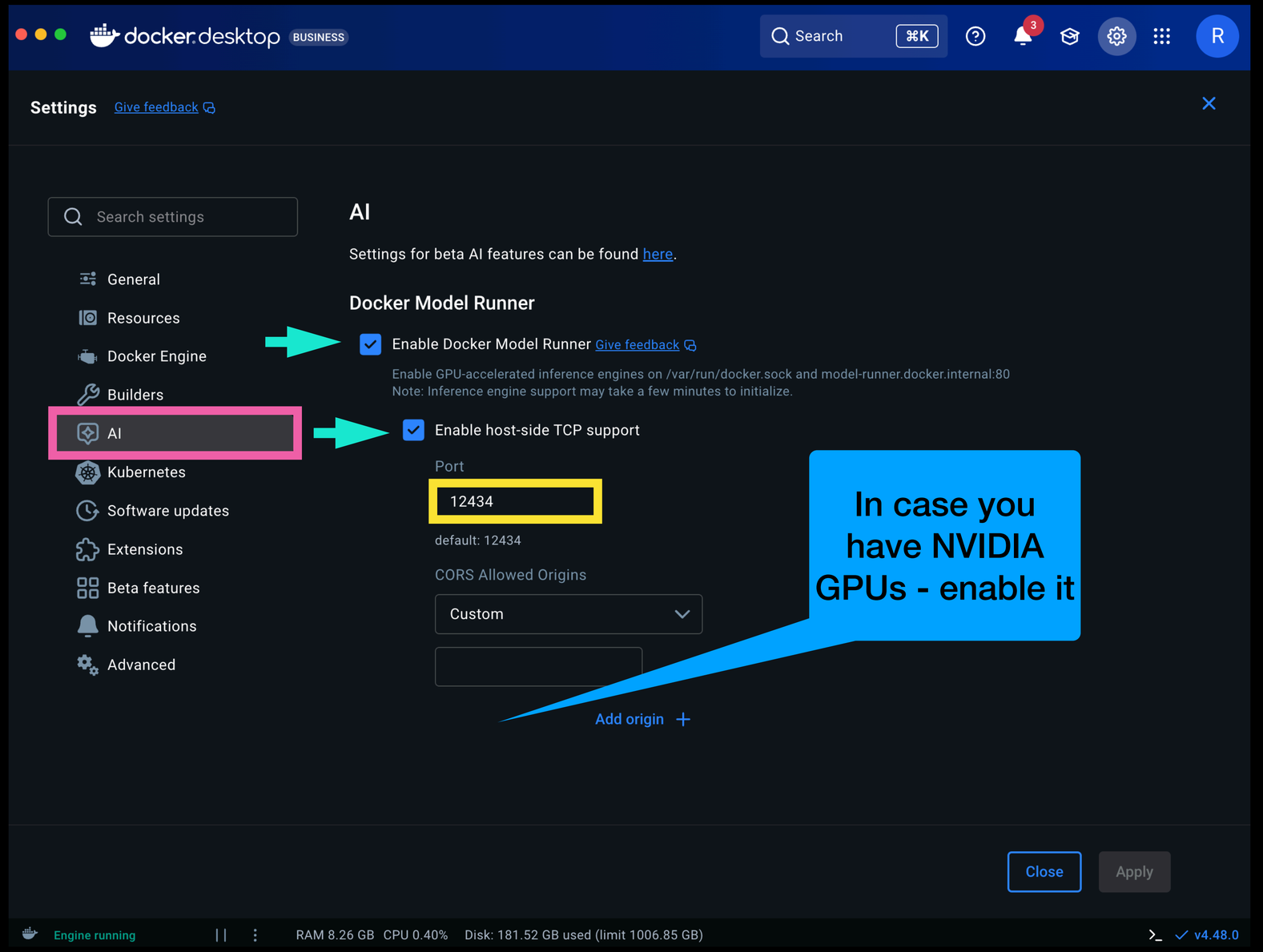
Using Paru on Arch Linux

```
paru -S opencode-bin
```

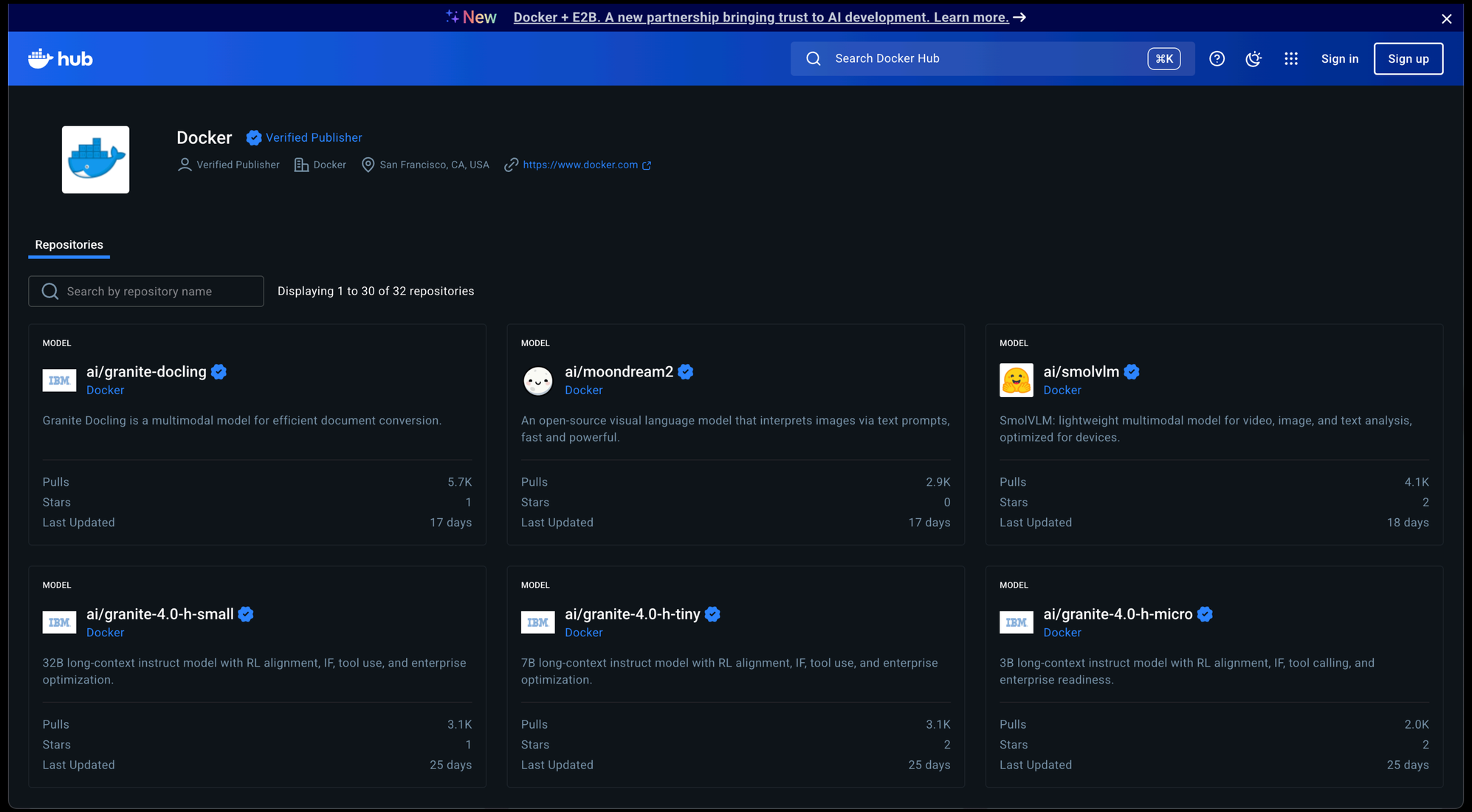
Setup


~
ramikrispin ~ ♡17:06 docker login
Authenticating with existing credentials... [Username: rkrispin]
i Info → To login with a different account, run 'docker logout' followed by 'docker login'
Login Succeeded
ramikrispin ~ ♡17:07 hf auth login --token $HF_TOKEN
The token has not been saved to the git credentials helper. Pass `add_to_git_credential=True` in this function directly or `--add-to-git-credential` if using via `hf`CLI if you want to set the git credential as well.
Token is valid (permission: fineGrained).
The token `DMR Workshop` has been saved to /Users/ramikrispin/.cache/huggingface/stored_tokens
Your token has been saved to /Users/ramikrispin/.cache/huggingface/token
Login successful.
Note: Environment variable`HF_TOKEN` is set and is the current active token independently from the token you've just configured.

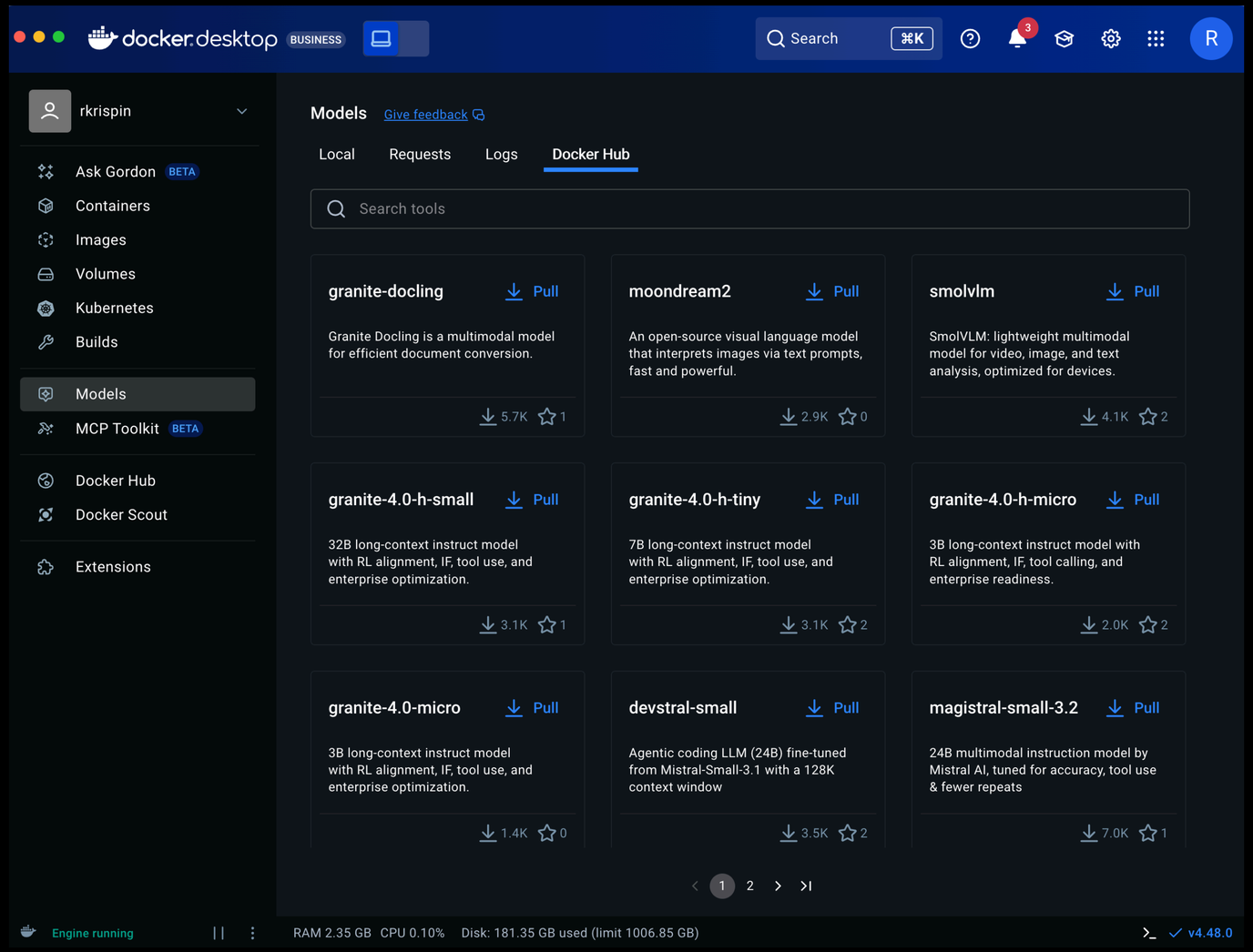
Setup


docker desktop
BUSINESS
Search
⌘K
3
R
Settings
Give feedback
Search settings
General
Resources
Docker Engine
Builders
AI
Kubernetes
Software updates
Extensions
Beta features
Notifications
Advanced
AI
Settings for beta AI features can be found here.
Docker Model Runner
Enable Docker Model Runner
Give feedback
Enable GPU-accelerated inference engines on /var/run/docker.sock and model-runner.docker.internal:80
Note: Inference engine support may take a few minutes to initialize.
Enable host-side TCP support
Port
12434
default: 12434
CORS Allowed Origins
Custom
Add origin
In case you have NVIDIA GPUs - enable it
Close
Apply
Engine running
RAM 8.26 GB CPU 0.40%
Disk: 181.52 GB used (limit 1006.85 GB)
v4.48.0

Models

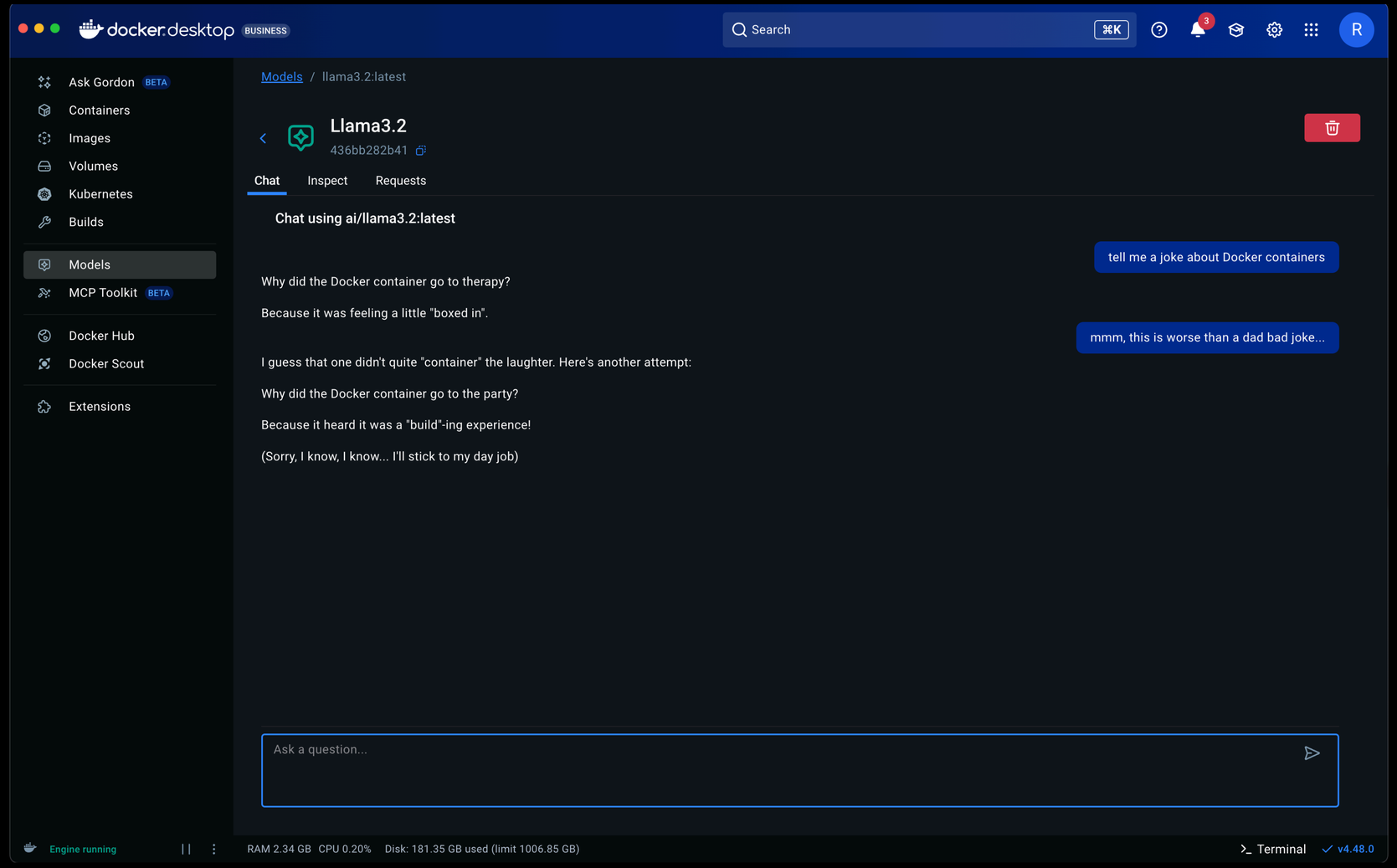

New
Docker + E2B. A new partnership bringing trust to AI development. Learn more.
hub
Search Docker Hub
⌘K
Sign in
Sign up
Docker
Verified Publisher
Verified Publisher
Docker
San Francisco, CA, USA
https://www.docker.com
Repositories
Search by repository name
Displaying 1 to 30 of 32 repositories
MODEL
ai/granite-docling
Docker
Granite Docling is a multimodal model for efficient document conversion.
Pulls 5.7K
Stars 1
Last Updated 17 days
MODEL
ai/moondream2
Docker
An open-source visual language model that interprets images via text prompts, fast and powerful.
Pulls 2.9K
Stars 0
Last Updated 17 days
MODEL
ai/smolvlm
Docker
SmolVLM: lightweight multimodal model for video, image, and text analysis, optimized for devices.
Pulls 4.1K
Stars 2
Last Updated 18 days
MODEL
ai/granite-4.0-h-small
Docker
32B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.
Pulls 3.1K
Stars 1
Last Updated 25 days
MODEL
ai/granite-4.0-h-tiny
Docker
7B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.
Pulls 3.1K
Stars 2
Last Updated 25 days
MODEL
ai/granite-4.0-h-micro
Docker
3B long-context instruct model with RL alignment, IF, tool calling, and enterprise readiness.
Pulls 2.0K
Stars 2
Last Updated 25 days

Models


docker desktop
BUSINESS
Search
⌘K
3
R
rkrispin
Ask Gordon
BETA
Containers
Images
Volumes
Kubernetes
Builds
Models
MCP Toolkit
BETA
Docker Hub
Docker Scout
Extensions
Models
Give feedback
Local
Requests
Logs
Docker Hub
Search tools
granite-docling
Pull
Granite Docling is a multimodal model for efficient document conversion.
5.7K
1
moondream2
Pull
An open-source visual language model that interprets images via text prompts, fast and powerful.
2.9K
0
smolvlm
Pull
SmolVLM: lightweight multimodal model for video, image, and text analysis, optimized for devices.
4.1K
2
granite-4.0-h-small
Pull
32B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.
3.1K
1
granite-4.0-h-tiny
Pull
7B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.
3.1K
2
granite-4.0-h-micro
Pull
3B long-context instruct model with RL alignment, IF, tool calling, and enterprise readiness.
2.0K
2
granite-4.0-micro
Pull
3B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.
1.4K
0
devstral-small
Pull
Agentic coding LLM (24B) fine-tuned from Mistral-Small-3.1 with a 128K context window
3.5K
2
magistral-small-3.2
Pull
24B multimodal instruction model by Mistral AI, tuned for accuracy, tool use & fewer repeats
7.0K
1
1
2
Engine running
RAM 2.35 GB CPU 0.10% Disk: 181.35 GB used (limit 1006.85 GB)
v4.48.0

Hello World!

Running LLMs via Docker Desktop



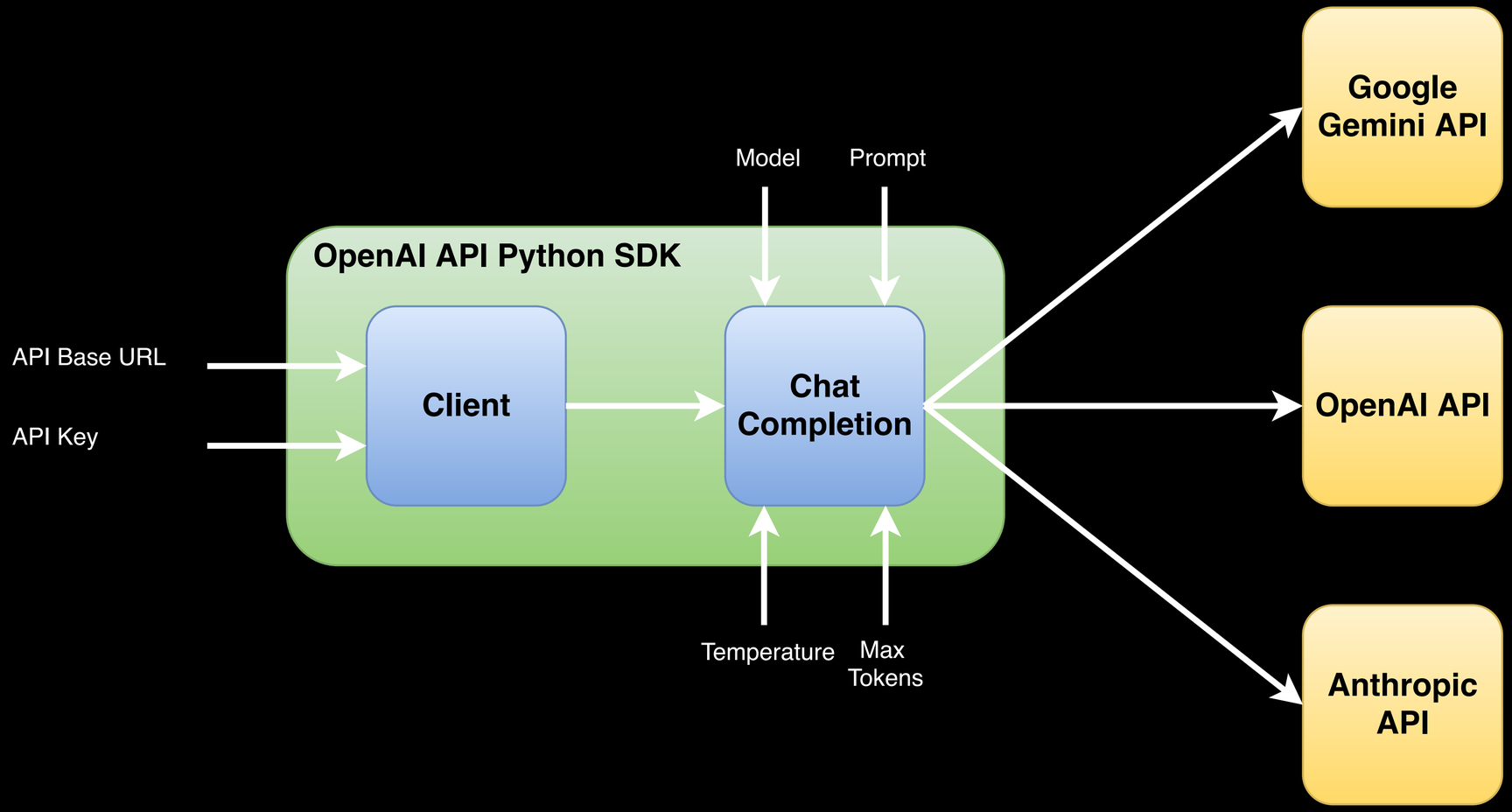
Running LLMs via the CLI

- CLI core commands
- Running LLMs
- Pull models from Docker Hub
- Pull LLMs from Hugging Face
- Update the context window

Running LLMs with Python

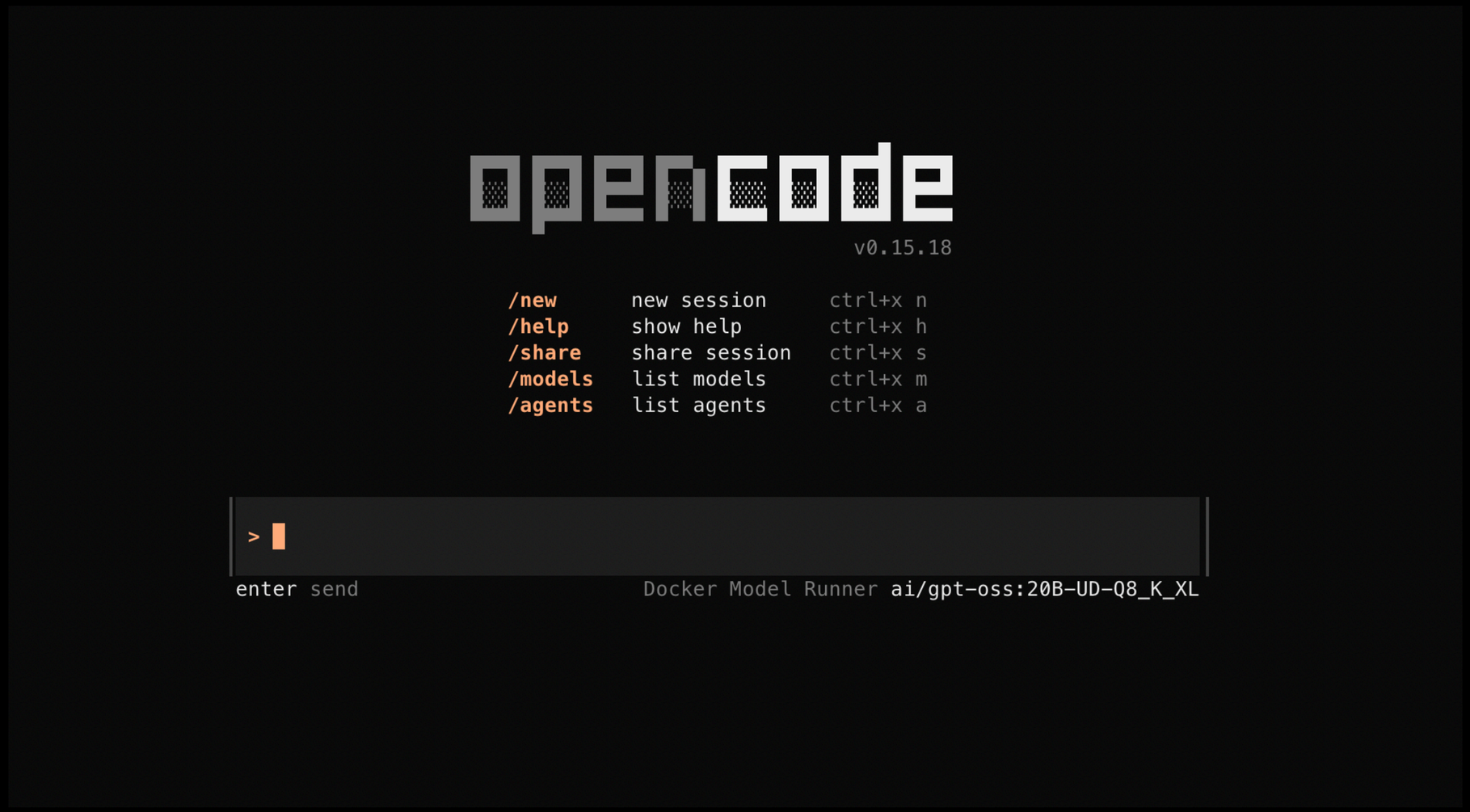
- OpenAI API SDK
- Calling LLMs from Virtual Environment vs. Container
- LangChain

OpenAI API SDK



Demo

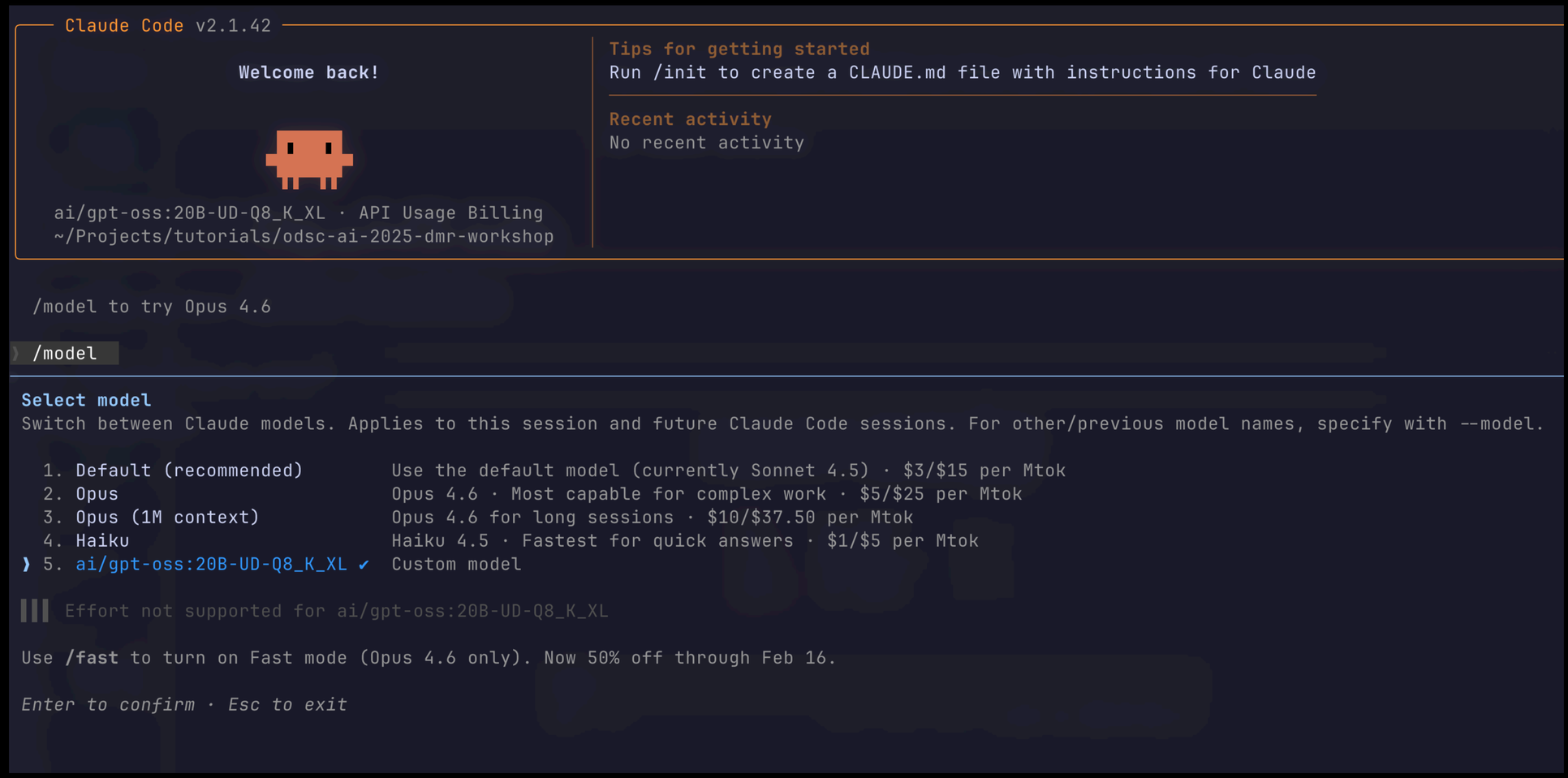
OpenCode


opencode
v0.15.18
/new new session ctrl+x n
/help show help ctrl+x h
/share share session ctrl+x s
/models list models ctrl+x m
/agents list agents ctrl+x a
>
enter send
Docker Model Runner ai/gpt-oss:20B-UD-Q8_K_XL

OpenCode

```
{
  "$schema": "https://opencode.ai/config.json",
  "provider": {
    "dmr": {
      "npm": "@ai-sdk/openai-compatible",
      "name": "Docker Model Runner",
      "options": {
        // Case using inside a container:
        // "baseURL": "http://model-runner.docker.internal/engines/v1",
        // Otherwise use the following base URL:
        "baseURL": "http://localhost:12434/engines/v1",
        "apiKey": "docker"
      },
      "models": {
        "ai/gpt-oss:20B-UD-Q8_K_XL": {
          "name": "ai/gpt-oss:20B-UD-Q8_K_XL"
        },
        "ai/llama3.2:3B-Q4_0": {
          "name": "ai/llama3.2:3B-Q4_0"
        },
        "ai/devstral-small:24B": {
          "name": "ai/devstral-small:24B"
        }
      }
    }
  }
}
```

Claude Code


Claude Code v2.1.42
Welcome back!
ai/gpt-oss:20B-UD-Q8_K_XL · API Usage Billing
~/Projects/tutorials/odsc-ai-2025-dmr-workshop
Tips for getting started
Run /init to create a CLAUDE.md file with instructions for Claude
Recent activity
No recent activity
/model to try Opus 4.6
❯ /model
Select model
Switch between Claude models. Applies to this session and future Claude Code sessions. For other/previous model names, specify with --model.
1. Default (recommended) Use the default model (currently Sonnet 4.5) · $3/$15 per Mtok
2. Opus Opus 4.6 · Most capable for complex work · $5/$25 per Mtok
3. Opus (1M context) Opus 4.6 for long sessions · $10/$37.50 per Mtok
4. Haiku Haiku 4.5 · Fastest for quick answers · $1/$5 per Mtok
❯ 5. ai/gpt-oss:20B-UD-Q8_K_XL ✔ Custom model
Effort not supported for ai/gpt-oss:20B-UD-Q8_K_XL
Use /fast to turn on Fast mode (Opus 4.6 only). Now 50% off through Feb 16.
Enter to confirm · Esc to exit

Claude Code

```
export ANTHROPIC_AUTH_TOKEN=docker
export ANTHROPIC_BASE_URL=http://localhost:12434
```

Claude Code

```
claude --model ai/gpt-oss:20B-UD-Q8_K_XL
```

Resources

- DMR documentation - https://docs.docker.com/ai/model-runner/
- Docker podcast - https://www.youtube.com/watch?v=9XryO1Oxb4A
- Workshop repo - https://github.com/RamiKrispin/dmr-workshop-packt
- The AIOps Newsletter - https://theaiops.substack.com/